

Canonical Correlation Analysis of Wine Data: A Multivariate Approach to Understanding Wine Characteristics

A Comparative Study of Canonical Correlation Approaches
presented by

RATHNAWEERA, R.P.R.M. (S/17/465)

DEPARTMENT OF STATISTICS & COMPUTER SCIENCE

FACULTY OF SCIENCE

UNIVERSITY OF PERADENIYA

SRI LANKA

2023

Table Of Contents

Introduction	3
Methodology	4
Results and Discussion	6
Conclusion and recommendation	11
References:	12
Appendices:	13

Introduction

Wine, a complex beverage with a rich history. This is a product of numerous factors that contribute to its final taste, aroma, and overall quality. These factors range from the type of grape used, the soil in which it was grown, the climate of the vineyard, to the fermentation process and the chemical composition of the wine itself. The study of wine is both fascinating and challenging due to the complex interaction of different factors. In this project, called "Canonical Correlation Analysis of Wine Data," we aim to explore and understand the connections between the chemical and sensory aspects of wine.

The dataset, which is used here, consists measurements of various chemical and sensory properties of wine. The chemical properties include factors such as alcohol content, acidity, residual sugar, and phenolic compounds, among others. On the other hand, the sensory properties encompass subjective assessments like taste, aroma, and colour. The primary objective of this project is to apply Canonical Correlation Analysis (CCA) to this dataset to identify and understand the relationships between these two sets of variables.

Canonical Correlation Analysis is a multivariate statistical technique that seeks to study the correlation between two sets of variables. In the context of this project, it will allow us to explore the associations between the chemical composition of wine and its sensory attributes.

In the following sections, I will present a detailed methodology of my analysis, followed by the results and their interpretation. I will conclude with a discussion on the implications of my findings.

Methodology

Dataset

The data used in this project Wine dataset, which can be found here:

<https://www.kaggle.com/datasets/harrywang/wine-dataset-for-clustering>

The dataset consists measurements of various chemical and sensory properties of wine. It has 179 observations and 13 columns. The columns containing following features,

- **Alcohol** is the percentage of alcohol by volume in the wine. It is a major determinant of the taste and strength of wine.
- **Malic acid** is a type of acid found in grapes. It gives wine its tartness.
- **Ash** is the inorganic residue that remains after wine grapes have been burned. It is a measure of the mineral content of the wine.
- **Alkalinity of ash** is a measure of the acidity of the ash. It is a measure of the wine's ability to buffer acids.
- **Magnesium** is a mineral that is found in wine grapes. It gives wine its bitterness.
- **Total phenols** are a group of compounds that are found in grape skins and seeds. They give wine its color, flavor, and aroma.
- **Flavanoids** are a type of phenol that is found in grape skins and seeds. They give wine its bitterness and astringency.
- **Nonflavanoid phenols** are a type of phenol that is found in grape skins and seeds. They give wine its bitterness and astringency.
- **Proanthocyanins** are a type of phenol that is found in grape skins and seeds. They give wine its color and astringency.
- **Color intensity** is a measure of the intensity of the wine's color. It is determined by the amount of anthocyanins and other pigments in the wine.
- **Hue** is a measure of the color of the wine. It is determined by the relative amounts of red, blue, and yellow pigments in the wine.
- **OD280/OD315 of diluted wines** is a measure of the absorbance of the wine at two different wavelengths. It is used to determine the concentration of certain compounds in the wine, such as proteins and anthocyanins.
- **Proline** is an amino acid that is found in wine. It gives wine its body and mouthfeel.

Data Pre-processing

Data Cleaning: Removing any incomplete or inconsistent records from the dataset.

Then, separate the 13 variables into two groups. The first group has **Chemical properties**. (Alcohol, Malic acid, Ash, Alkalinity of ash, Magnesium, Total phenols, Flavanoids, Nonflavanoid phenols, Proanthocyanins, OD280) and The second group has **Sensory properties** (Color intensity, Hue, Alkalinity of ash and Proline). Then I standardize the variables in dataset before applying Canonical Correlation Analysis (CCA), here what does is transform each variable to have a mean of 0 and a standard deviation of 1, ensuring that all variables are on the same scale.

Tests Used

a. Wilks' Lambda:

Wilks' Lambda is a statistical test used in multivariate analysis to assess the significance of the relationship between two sets of variables. It measures the proportion of variance in one set of variables that cannot be accounted for by the other set. Wilks' Lambda can help determine the overall significance of the relationship between the chemical properties and sensory properties of wine.

The null hypothesis: there is no significant relationship between the chemical properties and sensory properties of wine.

The alternative hypothesis: there is a significant relationship between the chemical properties and sensory properties of wine

Statistical Methods

The project uses statistical method as Canonical Correlation Analysis (CCA). Canonical Correlation Analysis is a multivariate statistical technique used to find the relationships between two sets of variables. It aims to identify linear combinations of variables from each set that are maximally correlated with each other. Canonical Correlation Analysis allows us to summarize the relationship into lesser number of statistics while preserving the main facets of the relationships. In this project, CCA allows us to examine the associations between the chemical properties and sensory properties of wine, providing a comprehensive understanding of their interdependencies.

Results and Discussion

Canonical Correlation Analysis

Split the dataset into two sets. Set one is “Chemical_properties” & set two is “Sensory_properties”. From this we can conclude that, there is only 4 canonical covariate pairs. (“Sensory_properties” set has four variables and “Chemical_properties” has only 9 variables). Then, To avoid problems in the computations, convert the data into the matrix format before performing CCA.

Chemical_properties

Alcohol <dbl>	Malic_Acid <dbl>	Ash <dbl>	Ash_Alcanity <dbl>	Magnesium <int>	Total_Phenols <dbl>	Flavanoids <dbl>	Nonflavanoid_Phenols <dbl>	Proanthocyanins <dbl>
14.23	1.71	2.43	15.6	127	2.80	3.06	0.28	2.29
13.20	1.78	2.14	11.2	100	2.65	2.76	0.26	1.28
13.16	2.36	2.67	18.6	101	2.80	3.24	0.30	2.81
14.37	1.95	2.50	16.8	113	3.85	3.49	0.24	2.18
13.24	2.59	2.87	21.0	118	2.80	2.69	0.39	1.82
14.20	1.76	2.45	15.2	112	3.27	3.39	0.34	1.97
14.39	1.87	2.45	14.6	96	2.50	2.52	0.30	1.98
14.06	2.15	2.61	17.6	121	2.60	2.51	0.31	1.25
14.83	1.64	2.17	14.0	97	2.80	2.98	0.29	1.98
13.86	1.35	2.27	16.0	98	2.98	3.15	0.22	1.85
14.10	2.16	2.30	18.0	105	2.95	3.32	0.22	2.38
14.12	1.48	2.32	16.8	95	2.20	2.43	0.26	1.57
13.75	1.73	2.41	16.0	89	2.60	2.76	0.29	1.81
14.75	1.73	2.39	11.4	91	3.10	3.69	0.43	2.81
14.38	1.87	2.38	12.0	102	3.30	3.64	0.29	2.96
13.63	1.81	2.70	17.2	112	2.85	2.91	0.30	1.46
14.30	1.92	2.72	20.0	120	2.80	3.14	0.33	1.97
13.83	1.57	2.62	20.0	115	2.95	3.40	0.40	1.72
14.19	1.59	2.48	16.5	108	3.30	3.93	0.32	1.86
13.64	3.10	2.56	15.2	116	2.70	3.03	0.17	1.66
14.06	1.63	2.28	16.0	126	3.00	3.17	0.24	2.10
12.93	3.80	2.65	18.6	102	2.41	2.41	0.25	1.98
13.71	1.86	2.36	16.6	101	2.61	2.88	0.27	1.69
12.85	1.60	2.52	17.8	95	2.48	2.37	0.26	1.46
13.50	1.81	2.61	20.0	96	2.53	2.61	0.28	1.66
13.05	2.05	3.22	25.0	124	2.63	2.68	0.47	1.92
13.39	1.77	2.62	16.1	93	2.85	2.94	0.34	1.45
13.30	1.72	2.14	17.0	94	2.40	2.19	0.27	1.35
13.87	1.90	2.80	19.4	107	2.95	2.97	0.37	1.76
14.02	1.68	2.21	16.0	96	2.65	2.33	0.26	1.98
13.73	1.50	2.70	22.5	101	3.00	3.25	0.29	2.38
13.58	1.66	2.36	19.1	106	2.86	3.19	0.22	1.95
13.68	1.83	2.36	17.2	104	2.42	2.69	0.42	1.97
13.76	1.53	2.70	19.5	132	2.95	2.74	0.50	1.35
13.51	1.80	2.65	19.0	110	2.35	2.53	0.29	1.54
13.48	1.81	2.41	20.5	100	2.70	2.98	0.26	1.86
13.28	1.64	2.84	15.5	110	2.60	2.68	0.34	1.36
13.05	1.65	2.55	18.0	98	2.45	2.43	0.29	1.44

Sensory_properties

Color_Intensity <dbl>	Hue <dbl>	OD280 <dbl>	Proline <int>
5.640000	1.040	3.92	1065
4.380000	1.050	3.40	1050
5.680000	1.030	3.17	1185
7.800000	0.860	3.45	1480
4.320000	1.040	2.93	735
6.750000	1.050	2.85	1450
5.250000	1.020	3.58	1290
5.050000	1.060	3.58	1295
5.200000	1.080	2.85	1045
7.220000	1.010	3.55	1045
5.750000	1.250	3.17	1510
5.000000	1.170	2.82	1280
5.600000	1.150	2.90	1320
5.400000	1.250	2.73	1150
7.500000	1.200	3.00	1547
7.300000	1.280	2.88	1310
6.200000	1.070	2.65	1280
6.600000	1.130	2.57	1130
8.700000	1.230	2.82	1680
5.100000	0.960	3.36	845
5.650000	1.090	3.71	780
4.500000	1.030	3.52	770
3.800000	1.110	4.00	1035
3.930000	1.090	3.63	1015
3.500000	1.120	3.62	815

The canonical correlation model was fitted to analyze the relationship between the variables in the 'sensory_properties' set and 'Chemical_properties' set. By applying this model, we obtained a total of four canonical correlations, which is equal to the number of variables in the 'sensory_properties' set. These canonical correlations provide valuable insights into the associations between the variables within the model.

```
{r}
cca_result <- cc(cp_mat, sp_mat)
cca_result$cor

[1] 0.8778781 0.7366933 0.4610869 0.2892986
```

In canonical correlation analysis, the canonical correlation coefficient measures the strength and direction of the relationship between the canonical variates from each set. The coefficients range from -1 to 1, where 1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 indicates no correlation.

Based on this results, it appears that there are four pairs of canonical variates, and their corresponding canonical correlation coefficients are 0.8778781, 0.7366933, 0.4610869, and 0.2892986.

The first pair of canonical variates has the highest correlation coefficient of 0.8778781, indicating a strong positive relationship between the two sets of variables. The second pair has a correlation coefficient of 0.7366933, which is slightly lower but still indicates a relatively strong positive relationship. The third and fourth pairs have correlation coefficients of 0.4610869 and 0.2892986, respectively, suggesting weaker relationships.

Test for independence between canonical variate pairs

```
{r}
wilks(cancor(cp_mat, sp_mat))
```

Test of H0: The canonical correlations in the current row and all that follow are zero

	CanR	LR	test stat	approx F	numDF	denDF	Pr(> F)	
1	0.87788	0.07566	17.0767	36	620.07	< 2.2e-16	***	
2	0.73669	0.32993	9.3538	24	482.05	< 2.2e-16	***	
3	0.46109	0.72150	4.2296	14	334.00	7.174e-07	***	
4	0.28930	0.91631	2.5575	6	168.00	0.0214	*	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The results of the Test for Independence between canonical variate pairs using Wilks' Lambda test indicate that there is a significant relationship between the canonical variate pairs. The p-values for all the tests are extremely small (less than 0.05), indicating strong evidence against the null hypothesis.

Squared canonical correlations.

```
{r}
squared_cc <- rho^2
squared_cc

[1] 0.77067001 0.54271701 0.21260109 0.08369371
```

Squared canonical correlations measure the strength of the relationship between sets of variables in canonical correlation analysis. Each squared canonical correlation represents the proportion of shared variance between the corresponding canonical variates. According to this result, We can conclude 77.06% of the variation in first canonical variable in “Chemical_properties” set is explained by the variation in first canonical variable of “Sensory_properties” set and the 54.27% of variation explained by second canonical variable as same. Others have relatively low values. Therefore only first one is high canonical correlation and implies that only the first canonical correlation is important.

The estimated canonical coefficients for the “Chemical_properties” set.

```
{r}
cca_result$xcoef
```

	[,1]	[,2]	[,3]	[,4]
Alcohol	-0.32062969	0.72898468	-0.02328237	0.5345369
Malic_Acid	0.14263586	0.06069861	-0.84800679	-0.5580144
Ash	-0.15238332	0.26950881	0.04151159	-0.2346341
Ash_Alcanity	0.15747380	-0.20360183	-0.37791152	0.9109756
Magnesium	-0.02608957	0.17300959	0.33018621	-0.3122841
Total_Phenols	-0.16938985	0.04525220	-0.43769660	0.3233158
Flavanoids	-0.46064434	-0.77132746	0.15907560	-0.2953254
Nonflavanoid_Phenols	0.06152172	0.06347417	0.51816718	0.3907676
Proanthocyanins	-0.07612897	0.14223507	-0.21080565	0.4166267

The magnitudes of the coefficients give the contribution of the individual variables to the corresponding canonical variable. “Chemical_properties” variables give less contribution to the first canonical variable of that set. Alcohol and Flavanoids give high contribution to second canonical variable of this set. Malic_Acid gives highest contribution to 3rd canonical variable and Ash_alcanity gives highest contribution to fourth canonical variable of this set.

The estimated canonical coefficients for the “Sensory_properties” set.

```
{r}
cca_result$ycoef
```

	[,1]	[,2]	[,3]	[,4]
Color_Intensity	-0.3612645	0.3974287	-0.1334638	1.3835898
Hue	-0.2316140	-0.2223699	1.0650956	0.8392436
OD280	-0.6573305	-0.4620249	-0.9944966	0.4044051
Proline	-0.4272639	0.5386522	0.2818859	-1.0882256

“Sensory_properties” variables give less contribution to the first and second canonical variables of that set. Hue and OD280 give highest contribution to third canonical variable and Color_Intensity and Proline give highest contribution to fourth canonical variable of this set.

The correlation between the “Chemical_properties” and the canonical variables for “Chemical_properties” set

```
{r}
loadings$corr.X.xscores
```

	[,1]	[,2]	[,3]	[,4]
Alcohol	-0.5733781	0.74170906	-0.08637612	0.084597183
Malic_Acid	0.4151581	0.39459526	-0.69077634	-0.230653559
Ash	-0.1986053	0.32325880	-0.11913603	0.185814731
Ash_Alcanity	0.4862455	-0.05577696	-0.31065996	0.565313721
Magnesium	-0.3377654	0.33735995	0.16820619	-0.270380756
Total_Phenols	-0.8582654	-0.23542205	-0.18741566	0.099135503
Flavanoids	-0.9024034	-0.38939785	-0.09042232	-0.006359933
Nonflavanoid_Phenols	0.5404541	0.24224895	0.24817323	0.370937852
Proanthocyanins	-0.6170607	-0.18677711	-0.22754457	0.219301063

Total_phenols and Flavanoids, Correlations are relatively large. There fore we can think these measures are overall measure of Chemical_properties. For the second, third, and fourth variables for chemical_properties, none of the correlations are particularly large, and so, this canonical variable yield little information about the data.

The correlation between the “Sensory_properties” and the canonical variables for “Chemical_properties” set

```
{r}
loadings$corr.Y.xscores
```

	[,1]	[,2]	[,3]	[,4]
Color_Intensity	-0.08216084	0.6496570	-0.08008411	0.12389453
Hue	-0.45273377	-0.4153426	0.29461571	0.02572603
OD280	-0.67334731	-0.4344442	-0.11380758	-0.01582114
Proline	-0.70383920	0.3442243	0.08309528	-0.09436203

Here we consider only first canonical variable because other canonical variables have small correlations. According to first canonical variable, OD280 and Proline correlations are moderately strong.

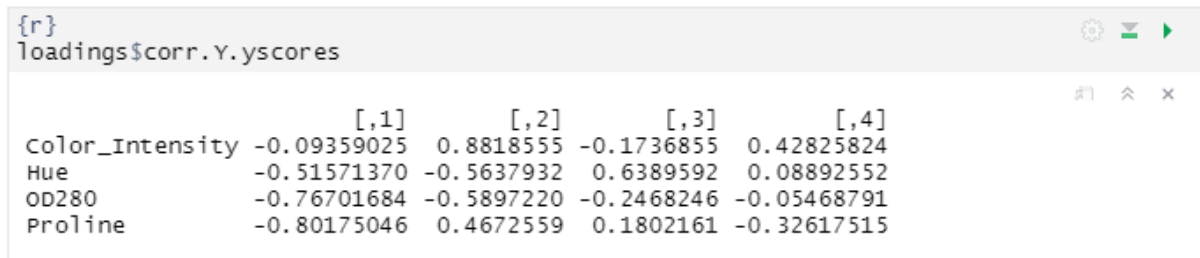
The correlation between the “Chemical_properties” and the canonical variables for “Sensory_properties” set

```
{r}
loadings$corr.X.yscores
```

	[,1]	[,2]	[,3]	[,4]
Alcohol	-0.5033561	0.54641209	-0.03982689	0.02447385
Malic_Acid	0.3644582	0.29069569	-0.31850789	-0.06672776
Ash	-0.1743513	0.23814259	-0.05493206	0.05375595
Ash_Alcanity	0.4268643	-0.04109051	-0.14324123	0.16354449
Magnesium	-0.2965169	0.24853081	0.07755766	-0.07822079
Total_Phenols	-0.7534525	-0.17343385	-0.08641490	0.02867977
Flavanoids	-0.7922002	-0.28686678	-0.04169254	-0.00183992
Nonflavanoid_Phenols	0.4744528	0.17846318	0.11442942	0.10731182
Proanthocyanins	-0.5417041	-0.13759744	-0.10491781	0.06344350

Total_Phenols and Flavanoids are moderately correlated with first canonical variable in this set and we only consider this. Because other canonical variables have relatively small correlations. So there is no much effect from those.

The correlation between the “Sensory_properties” and the canonical variables for “Sensory_properties” set



The image shows a screenshot of an R console window. The title bar is grey and contains the text "{r}" on the left and three icons (a gear, a green bar, and a green arrow) on the right. The main area of the console is white and displays the command "loadings\$corr.Y.ycores" followed by a matrix of correlation coefficients. The matrix has four columns labeled "[,1]", "[,2]", "[,3]", and "[,4]" at the top. The rows are labeled "Color_Intensity", "Hue", "OD280", and "Proline" on the left. The values are as follows:

	[,1]	[,2]	[,3]	[,4]
Color_Intensity	-0.09359025	0.8818555	-0.1736855	0.42825824
Hue	-0.51571370	-0.5637932	0.6389592	0.08892552
OD280	-0.76701684	-0.5897220	-0.2468246	-0.05468791
Proline	-0.80175046	0.4672559	0.1802161	-0.32617515

OD280 and Proline , highly correlated with first canonical variable and this show similar pattern to that with the first canonical variate for “Chemical_properties”.

Conclusion and recommendation

Conclusion

From the analysis we can conclude that don't want to get 36 pairwise scatterplots to explain the dataset. We can do it from only four canonical variate pairs. I proved it by "Wilki's lambda" test.

From squared canonical correlation we can conclude, 77.06 % of the variation in first canonical variable of "Chemical_properties" set is explained by the variation in first canonical variable of "Sensory_Properties" set.

This finding suggests that the sensory characteristics of wine, such as taste, aroma, and overall quality, are strongly influenced by the levels of OD280 and Proline. On the other hand, the chemical composition of wine, including the presence of total phenols and flavonoids, plays a significant role in determining its chemical properties.

Recommendation

Based on these results, it is recommended that winemakers to focus on monitoring and manipulating the levels of OD280, Proline, total phenols, and flavonoids to enhance both the sensory and chemical aspects of wine. By understanding the relationship between these variables, it becomes possible to optimize wine production processes, improve wine quality, and meet consumer preferences.

References:

(*Canonical Correlation Analysis*, n.d.)

(*Canonical Correlation Analysis | R Data Analysis Examples*, n.d.)

('Using the Canonical Correlation Analysis Method to Study Students' Levels in Face-to-Face and Online Education in Jordan', 2023)

Korstanje, J. (n.d.). Canonical Correlation Analysis.

Appendices:

Dataset

	A	B	C	D	E	F	G	H	I	J	K	L	M	
1	Alcohol	Malic_Acid	Ash	Ash_Alcal	Magnesium	Total_Phenols	Flavanoid	Nonflavonoid	Proanthocyanins	Color_Intensity	Hue	OD280	Proline	
2	14.23	1.71	2.43	15.6	127	2.8	3.06	0.28	2.29	5.64	1.04	3.92	1065	
3	13.2	1.78	2.14	11.2	100	2.65	2.76	0.26	1.28	4.38	1.05	3.4	1050	
4	13.16	2.36	2.67	18.6	101	2.8	3.24	0.3	2.81	5.68	1.03	3.17	1185	
5	14.37	1.95	2.5	16.8	113	3.85	3.49	0.24	2.18	7.8	0.86	3.45	1480	
5	13.24	2.59	2.87	21	118	2.8	2.69	0.39	1.82	4.32	1.04	2.93	735	
7	14.2	1.76	2.45	15.2	112	3.27	3.39	0.34	1.97	6.75	1.05	2.85	1450	
8	14.39	1.87	2.45	14.6	96	2.5	2.52	0.3	1.98	5.25	1.02	3.58	1290	
9	14.06	2.15	2.61	17.6	121	2.6	2.51	0.31	1.25	5.05	1.06	3.58	1295	
0	14.83	1.64	2.17	14	97	2.8	2.98	0.29	1.98	5.2	1.08	2.85	1045	
1	13.86	1.35	2.27	16	98	2.98	3.15	0.22	1.85	7.22	1.01	3.55	1045	
2	14.1	2.16	2.3	18	105	2.95	3.32	0.22	2.38	5.75	1.25	3.17	1510	
3	14.12	1.48	2.32	16.8	95	2.2	2.43	0.26	1.57	5	1.17	2.82	1280	
4	13.75	1.73	2.41	16	89	2.6	2.76	0.29	1.81	5.6	1.15	2.9	1320	
5	14.75	1.73	2.39	11.4	91	3.1	3.69	0.43	2.81	5.4	1.25	2.73	1150	
6	14.38	1.87	2.38	12	102	3.3	3.64	0.29	2.96	7.5	1.2	3	1547	
7	13.63	1.81	2.7	17.2	112	2.85	2.91	0.3	1.46	7.3	1.28	2.88	1310	
8	14.3	1.92	2.72	20	120	2.8	3.14	0.33	1.97	6.2	1.07	2.65	1280	
9	13.83	1.57	2.62	20	115	2.95	3.4	0.4	1.72	6.6	1.13	2.57	1130	
0	14.19	1.59	2.48	16.5	108	3.3	3.93	0.32	1.86	8.7	1.23	2.82	1680	
1	13.64	3.1	2.56	15.2	116	2.7	3.03	0.17	1.66	5.1	0.96	3.36	845	
2	14.06	1.63	2.28	16	126	3	3.17	0.24	2.1	5.65	1.09	3.71	780	
3	12.93	3.8	2.65	18.6	102	2.41	2.41	0.25	1.98	4.5	1.03	3.52	770	
4	13.71	1.86	2.36	16.6	101	2.61	2.88	0.27	1.69	3.8	1.11	4	1035	
5	12.85	1.6	2.52	17.8	95	2.48	2.37	0.26	1.46	3.93	1.09	3.63	1015	
6	13.5	1.81	2.61	20	96	2.53	2.61	0.28	1.66	3.52	1.12	3.82	845	
7	13.05	2.05	3.22	25	124	2.63	2.68	0.47	1.92	3.58	1.13	3.2	830	
8	13.39	1.77	2.62	16.1	93	2.85	2.94	0.34	1.45	4.8	0.92	3.22	1195	
9	13.3	1.72	2.14	17	94	2.4	2.19	0.27	1.35	3.95	1.02	2.77	1285	
0	13.87	1.9	2.8	19.4	107	2.95	2.97	0.37	1.76	4.5	1.25	3.4	915	
1	14.02	1.68	2.21	16	96	2.65	2.33	0.26	1.98	4.7	1.04	3.59	1035	
2	13.73	1.5	2.7	22.5	101	3	3.25	0.29	2.38	5.7	1.19	2.71	1285	
3	13.58	1.66	2.36	19.1	106	2.86	3.19	0.22	1.95	6.9	1.09	2.88	1515	
4	13.68	1.83	2.36	17.2	104	2.42	2.69	0.42	1.97	3.84	1.23	2.87	990	

Codes

```
library(readr)
library(CCA)
library(CCP)
library(candisc)

df <- read.csv('E:/Uni Docs/4th year/1st Sem/ST405/Mini Project 2/wine.csv')

library(dplyr)

# Check for null values in each column
missing_values <- df %>%
  summarise_all(~ sum(is.na(.)))

# Print the number of missing values in each column
print(missing_values)
```

```
##   Alcohol Malic_Acid Ash Ash_Alcanity Magnesium Total_Phenols Flavanoid
s
## 1      0      0  0      0      0      0
0
##   Nonflavanoid_Phenols Proanthocyanins Color_Intensity Hue OD280 Proline
## 1      0      0      0  0  0
0

colnames(df)

## [1] "Alcohol"           "Malic_Acid"         "Ash"
## [4] "Ash_Alcanity"       "Magnesium"          "Total_Phenols"
## [7] "Flavanoids"         "Nonflavanoid_Phenols" "Proanthocyanins"
## [10] "Color_Intensity"    "Hue"                "OD280"
## [13] "Proline"
```

Separate the 13 variables into two groups. The first group has 9 variables Alcohol, Malic_Acid, Ash, Ash_Alcanity, Magnesium, Total_Phenols, Flavanoids, Nonflavanoid_Phenols, Proanthocyanins. The second group has 4 variables Color_Intensity, Hue, OD280 and Proline.

```
chemical_properties <- cbind(df[,1:9])
sensory_properties <- cbind(df[,10:13])
```

Standardize all variables and apply CCA . This function will transform each variable to have a mean of 0 and a standard deviation of 1, ensuring that all variables are on the same scale.

```
library(dplyr)

chemical_properties_st <- chemical_properties %>%
  mutate(across(everything(), scale))

sensory_properties_st <- sensory_properties %>%
  mutate(across(everything(), scale))

Convert to matrix form

cp_mat <- as.matrix(chemical_properties_st)
sp_mat <- as.matrix(sensory_properties_st)

cca_result <- cc(cp_mat, sp_mat)
cca_result$cor

## [1] 0.8778781 0.7366933 0.4610869 0.2892986

can_variates_cp <- cca_result$xcoef
can_variates_sp <- cca_result$ycoef
rho <- cca_result$cor

n <- dim(chemical_properties)[1]
p <- dim(chemical_properties)[2]
q <- dim(sensory_properties)[2]
```

```

print(n)
## [1] 178

print(p)
## [1] 9

print(q)
## [1] 4

print(rho)
## [1] 0.8778781 0.7366933 0.4610869 0.2892986

library(CCP)

p.asym(rho,n,p,q,tstat = "Wilks")

## Wilks' Lambda, using F-approximation (Rao's F):
##
##      stat      approx df1      df2      p.value
## 1 to 4: 0.07566262 17.076703 36 620.0687 0.000000e+00
## 2 to 4: 0.32992902 9.353770 24 482.0517 0.000000e+00
## 3 to 4: 0.72149858 4.229556 14 334.0000 7.173509e-07
## 4 to 4: 0.91630629 2.557468 6 168.0000 2.140004e-02

p.asym(rho,n,p,q,tstat = "Hotelling")

## Hotelling-Lawley Trace, using F-approximation:
##
##      stat      approx df1 df2      p.value
## 1 to 4: 4.90870071 22.293682 36 654 0.000000e+00
## 2 to 4: 1.54817209 10.675937 24 662 0.000000e+00
## 3 to 4: 0.36134243 4.323204 14 670 2.181344e-07
## 4 to 4: 0.09133813 2.580302 6 678 1.768376e-02

p.asym(rho,n,p,q,tstat = "Pillai")

## Pillai-Bartlett Trace, using F-approximation:
##
##      stat      approx df1 df2      p.value
## 1 to 4: 1.60968182 12.570458 36 672 0.000000e+00
## 2 to 4: 0.83901181 7.520433 24 680 0.000000e+00
## 3 to 4: 0.29629480 3.931407 14 688 1.621931e-06
## 4 to 4: 0.08369371 2.478986 6 696 2.221297e-02

p.asym(rho,n,p,q,tstat = "Roy")

## Roy's Largest Root, using F-approximation:
##
##      stat      approx df1 df2      p.value
## 1 to 1: 0.77067 145.3429 4 173 0
##
## F statistic for Roy's Greatest Root is an upper bound.

Wilks(cancor(cp_mat,sp_mat))

##
## Test of H0: The canonical correlations in the

```

```

## current row and all that follow are zero
##
##      CanR LR test stat approx F numDF  denDF    Pr(> F)
## 1 0.87788      0.07566 17.0767    36 620.07 < 2.2e-16 ***
## 2 0.73669      0.32993  9.3538    24 482.05 < 2.2e-16 ***
## 3 0.46109      0.72150  4.2296    14 334.00 7.174e-07 ***
## 4 0.28930      0.91631  2.5575     6 168.00  0.0214 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

squared_cc <- rho^2
squared_cc

## [1] 0.77067001 0.54271701 0.21260109 0.08369371

cca_result$xcoef

##              [,1]      [,2]      [,3]      [,4]
## Alcohol        -0.32062969  0.72898468 -0.02328237  0.5345369
## Malic_Acid      0.14263586  0.06069861 -0.84800679 -0.5580144
## Ash            -0.15238332  0.26950881  0.04151159 -0.2346341
## Ash_Alcanity    0.15747380 -0.20360183 -0.37791152  0.9109756
## Magnesium       -0.02608957  0.17300959  0.33018621 -0.3122841
## Total_Phenols   -0.16938985  0.04525220 -0.43769660  0.3233158
## Flavanoids      -0.46064434 -0.77132746  0.15907560 -0.2953254
## Nonflavanoid_Phenols 0.06152172  0.06347417  0.51816718  0.3907676
## Proanthocyanins -0.07612897  0.14223507 -0.21080565  0.4166267

cca_result$ycoef

##              [,1]      [,2]      [,3]      [,4]
## Color_Intensity -0.3612645  0.3974287 -0.1334638  1.3835898
## Hue             -0.2316140 -0.2223699  1.0650956  0.8392436
## OD280           -0.6573305 -0.4620249 -0.9944966  0.4044051
## Proline         -0.4272639  0.5386522  0.2818859 -1.0882256

loadings <- comput(cp_mat,sp_mat,cca_result)

loadings$corr.X.xscores

##              [,1]      [,2]      [,3]      [,4]
## Alcohol        -0.5733781  0.74170906 -0.08637612  0.084597183
## Malic_Acid      0.4151581  0.39459526 -0.69077634 -0.230653559
## Ash            -0.1986053  0.32325880 -0.11913603  0.185814731
## Ash_Alcanity    0.4862455 -0.05577696 -0.31065996  0.565313721
## Magnesium       -0.3377654  0.33735995  0.16820619 -0.270380756
## Total_Phenols   -0.8582654 -0.23542205 -0.18741566  0.099135503
## Flavanoids      -0.9024034 -0.38939785 -0.09042232 -0.006359933
## Nonflavanoid_Phenols 0.5404541  0.24224895  0.24817323  0.370937852
## Proanthocyanins -0.6170607 -0.18677711 -0.22754457  0.219301063

loadings$corr.Y.xscores

```


##		[,1]	[,2]	[,3]	[,4]
## Color_Intensity		-0.08216084	0.6496570	-0.08008411	0.12389453
## Hue		-0.45273377	-0.4153426	0.29461571	0.02572603
## OD280		-0.67334731	-0.4344442	-0.11380758	-0.01582114
## Proline		-0.70383920	0.3442243	0.08309528	-0.09436203

loadings\$corr.X.yscores

##		[,1]	[,2]	[,3]	[,4]
## Alcohol		-0.5033561	0.54641209	-0.03982689	0.02447385
## Malic_Acid		0.3644582	0.29069569	-0.31850789	-0.06672776
## Ash		-0.1743513	0.23814259	-0.05493206	0.05375595
## Ash_Alcanity		0.4268643	-0.04109051	-0.14324123	0.16354449
## Magnesium		-0.2965169	0.24853081	0.07755766	-0.07822079
## Total_Phenols		-0.7534525	-0.17343385	-0.08641490	0.02867977
## Flavanoids		-0.7922002	-0.28686678	-0.04169254	-0.00183992
## Nonflavanoid_Phenols		0.4744528	0.17846318	0.11442942	0.10731182
## Proanthocyanins		-0.5417041	-0.13759744	-0.10491781	0.06344350

loadings\$corr.Y.yscores

##		[,1]	[,2]	[,3]	[,4]
## Color_Intensity		-0.09359025	0.8818555	-0.1736855	0.42825824
## Hue		-0.51571370	-0.5637932	0.6389592	0.08892552
## OD280		-0.76701684	-0.5897220	-0.2468246	-0.05468791
## Proline		-0.80175046	0.4672559	0.1802161	-0.32617515